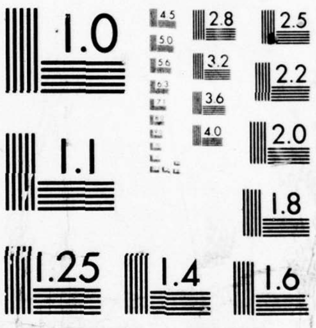
AD-A054 551 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
ANALYSIS OF EVOLUTIONARY ERROR IN FINITE ELEMENT AND OTHER METH--ETC(U)
FEB 78 M J CULLEN, K W MORTON DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1832 NL

| OF |

AD
A054551

END
DATE
FILMED
6-78
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

13

MRC Technical Summary Report # 1832

ANALYSIS OF EVOLUTIONARY ERROR IN FINITE ELEMENT AND OTHER METHODS

M. J. P. Cullen and K. W. Morton

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

February 1978

(Received September 2, 1977)

Sponsored by

U.S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

Meteorological Office
Bracknell
England

University of Reading
Department of Mathematics
England

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

ANALYSIS OF EVOLUTIONARY ERROR IN FINITE ELEMENT AND OTHER METHODS

M. J. P. Cullen[1] and K. W. Morton[2]

Technical Summary Report #1832

February 1978

ABSTRACT

Restriction and prolongation operators are used to provide a unified framework for the discussion of errors in approximating evolutionary equations. A generalized truncation error enables the spline-Galerkin method to be studied in detail and the accuracy of various treatments of non-linear terms (such as the advection operator $\underline{v}\cdot\nabla\underline{v}$) compared: it is shown how a multi-stage Galerkin process can give errors which are $O(h^{2\mu})$ for splines of order μ and quite general differential operators. A Petrov-Galerkin method is derived for $\partial_t = a\partial_x u$ which is accurate and stable.

AMS(MOS) Subject Classifications: 35A40, 41A15, 65M05, 65M10, 65M15, 65N30.

Key Words: Error analysis, Finite difference methods, Finite element methods, Spline-Galerkin, Petrov-Galerkin, Prolongation and restriction operators, Superconvergence, Advection equation.

Work Unit Number 7 - Numerical Analysis

See 1473

[1] Supported by the Meteorological Office, in partial fulfillment of the requirements for a Ph.D. at University of Reading.

[2] Sponsored in part by the United States Army under Contract No. DAAG29-75-C-0024 and in part by the University of Reading Mathematics Department.

SIGNIFICANCE AND EXPANATION

Most problems in continuum mechanics (fluid flow, combustion, elasticity) involve partial differential equations that can be solved only numerically by computer. Originally most numerical methods for this class of problem used finite differences, which involve direct replacement of derivatives by difference formulae. In the last ten or fifteen years, finite element methods have become increasingly popular. These involve starting from an assumed functional form for the unknowns (e.g. piecewise polynomial), then determining parameters in the functional representation via satisfying the partial differential equation in some approximate sense.

As finite element methods become increasingly used for non-steady problems, it becomes important that their performance can be compared in detail with that of the longer established finite difference methods. This paper sets up a framework in which this can be done. Analysis of the spline-Galerkin methods is carried out and highly accurate schemes put forward for the advection operator $\underline{v}\cdot\underline{\nabla}\,\underline{v}$, which occurs in the equations for practical problems in which motion of fluid and material particles are involved.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

ANALYSIS OF EVOLUTIONARY ERROR IN FINITE ELEMENT AND OTHER METHODS

M. J. P. Cullen[1] and K. W. Morton[2]

1. INTRODUCTION

Any step-by-step procedure for approximating an initial-value problem $u_t = Lu$ or $u_{tt} = Lu$ consists of three stages: discretisation of the initial data; updating the discrete approximant to match the evolution of the true solution; and interpretation of the final result. In this paper, we present a general framework for describing and analysing such procedures based on the restriction and prolongation operators introduced by Aubin [1] for studying elliptic problems and used by Noble [20] and others for Fredholm integral equations. The first stage of a procedure naturally entails the use of a restriction operator r_h, the second an evolution operator E_h and the last a prolongation p_h. Our objective is a practical one: to provide a framework within which in particular finite element, finite difference and spectral methods can be compared in detail and the development of hybrid methods encouraged. Detailed comparisons mainly refer to the advection operator $\underline{v}\cdot\underline{\nabla}\,\underline{v}$

Following Swartz and Wendroff [28], Douglas and Dupont [8] and Dupont [10], the standard error analysis of Galerkin methods when Lu is linear, elliptic and negative definite introduces an elliptic projection of the solution Pu and estimates the rate of evolution of the difference $Pu - U$, where U is the Galerkin approximation. Much of the theory for elliptic problems can then be invoked to show that this "evolutionary error" has the optimal rate of convergence $O(h^k)$ in the L^2 norm, where h is a space discretisation parameter and the approximation space or trial space has order of accuracy h^k. The analysis is based on energy estimates and is particularly appropriate when L is self-adjoint. For then the Ritz projection Pu is in a definite sense the best approximation to u and gives a firm basis of comparison for U. But the approach can also be used for more general elliptic operators (see Douglas, Dupont and Wheeler [9]) and for some non-linear problems (see, for instance, Wheeler [34] and Dendy [7]).

[1] Supported by the Meteorological Office, in partial fulfillment of the requirements for a Ph.D. at University of Reading.

[2] Sponsored in part by the United States Army under Contract No. DAAG29-75-C-0024 and by the University of Reading Mathematics Department.

However, as the departure from self-adjointness is increased, Pu becomes a less and less good approximation to u or, if a weighted projection is used, the penalty is paid in an extra growth term for $Pu - U$. Coercivity of L is enough to obtain optimal order of accuracy but the constants may be large. Indeed, at the extreme of first order hyperbolic equations even this may be lost, as was shown by the example of Dupont [11] - even though Dendy [6] demonstrated how this could be recovered. Nevertheless, the general approach is still valid in all cases: U should be compared with a projection of u into the approximation space and much information is lost if it is compared directly with u. In integrating over any reasonable length of time, the evolutionary error is dominant and the most useful schemes will often exhibit phenomena of superconvergence which can best be studied through such comparisons. Unfortunately the standard analysis precludes this by the presence of a term $P\partial_t u - \partial_t U$. We present an alternative analysis which places more emphasis on the discrete procedure: it is quite general both as regard to type of procedure and the projection used in the comparison and, by means of a generalised truncation error, allows superconvergence of the evolutionary error to be easily studied.

A particularly interesting case of superconvergence was exhibited by Thomée [32] and Thomée and Wendroff [33]. They showed that linear elements used with a Galerkin procedure on first order linear hyperbolic equations could be interpreted as giving a finite difference scheme with $O(h^4)$ accuracy, and more generally splines of order μ gave an accuracy of $O(h^{2\mu})$. We shall show that this is true under any projection and that with orthogonal protion in L^2 the superconvergence continues to hold with non-linear terms like $v_x u$. This view-point draws attention to the possibility of evaluating such a product in a number of different ways - pointwise multiplication, simple Galerkin and two-stage Galerkin - which have very different error characteristics. In numerical experiments on the shallow water equations, Cullen [3,4] found that the non-conservative two-stage Galerkin process gave very much better results than the single stage process: the analysis indicates why this is so. Similarly, higher order differential operators led to a loss of accuracy in Thomée's analysis - for instance, only $O(h^2)$ for the heat equation and linear elements: we shall show that a multistage process can retain the full $O(h^{2\mu})$ accuracy for any order, at the cost of a less compact scheme.

The plan of the rest of the paper is as follows. In section 2, restriction and prolongation operators are introduced and the unified framework for analysing evolutionary error is developed. This is illustrated in section 3 on a simple finite element problem and applied in section 4 to the semi-discrete spline Galerkin method. In both sections Fourier analysis is the main tool for estimating the errors. Section 5 places finite difference methods in the same framework, drawing attention to the optimal recovery problem arising at the final prolongation. A Petrov-Galerkin method for $u_t = au_x$ is proposed in section 6 which is motivated by reference to characteristics, shown to be stable by an energy analysis and its accuracy determined by Fourier analysis. Finally, in section 7, the results as they pertain to the advection operator are drawn together, the importance of the restriction operator in damping short-wave length modes emphasised and the accuracy of an irregular mesh considered.

2. UNIFIED ANALYSIS OF EVOLUTIONARY ERROR

We consider pure initial-value problems in the form

$$\partial_t u - Lu = f \quad \text{on} \quad [0,T] \times \mathbb{R}^d$$
$$u(0,x) = u^0(x) \quad , \tag{2.1}$$

where u is vector-valued (and may be complex in Fourier analysis) and L is a differential operator in $\mathbb{R}^d$, which may be non-linear but has real coefficients and, like f, does not depend explicitly on t. It is most convenient to work in a separable Hilbert space V with norm $\|\cdot\|$ and inner product $<\cdot,\cdot>$ and denote by $u(t)$ the mapping $u:[0,T] \to V$. We assume that L generates a strong evolution operator $E(\tau):V \to V$, $\tau \geq 0$, so that we may write

$$u(t+\tau) = E(\tau)u(t) \ , \ \tau \geq 0 \ . \tag{2.2}$$

(See, e.g., Kato [16] for conditions under which this may be established and Tartar [30] for a brief introduction.)

Following Aubin [1], we associate the triplet (V_h, p_h, r_h) with any procedure for approximating members of the solution space V on a discrete mesh in $\mathbb{R}^d$ which is characterised by a parameter $h > 0$. The discrete space V_h, with norm $\|\cdot\|_h$, consists of elements u_h which are finite dimensional vectors of parameters defining an approximation to u; the restriction operator $r_h:V \to V_h$ is a continuous mapping identifying this relationship; and the prolongation operator p_h is an isomorphism from V_h to a closed subspace S_h of V, called the approximation space. In a typical situation, S_h might consist of continuous piecewise linear functions over a triangulation of $\mathbb{R}^2$, u_h the set of nodal values at the vertices, $r_h u$ those values obtained from a least squares fit to u and $p_h r_h u$ the resultant linear approximation. For a given prolongation p_h, we can define the optimal restriction (in V), denoted by $\tilde{r}_h$, such that

$$\|u - p_h\tilde{r}_h u\| = \inf_{v_h \in V_h} \|u - p_h v_h\| \ , \ \forall u \in V \ . \tag{2.3}$$

It follows that

$$\tilde{r}_h p_h = I_h \ , \ \|p_h \tilde{r}_h\| = 1 \ , \tag{2.4}$$

where I_h is the identity in V_h. In fact we shall always ensure that $r_h p_h = I_h$, so that $p_h r_h$ is a projection, and $p_h \tilde{r}_h$ is the orthogonal projector: also the usual discrete norm is given by $\|u_h\|_h \equiv \|p_h u_h\|$, and we shall always assume that $\|r_h\|_h$ is uniformly bounded as $h \to 0$. Similarly for a given r_h, $\tilde{p}_h$ is called an <u>optimal prolongation</u> if

$$r_h \tilde{p}_h = I_h \quad \text{and} \quad \|\tilde{p}_h v_h\| = \inf_{\{v \in V | r_h v = v_h\}} \|v\| \ , \ \forall v_h \in V_h \ . \tag{2.5}$$

Part of the convenience of working in a Hilbert space is that a unique $\tilde{p}_h$ exists for each r_h and satisfies a dual property

$$\|\kappa - \tilde{p}_h r_h u\| = \inf_{v_h \in V_h} \|u - \tilde{p}_h v_h\| \ , \ \forall u \in V \ . \tag{2.6}$$

A <u>semi-discrete</u> approximation to (2.1) consists of a one parameter family $u_h(t)$, that is a mapping $u_h:[0,T] \to V_h$, satisfying a system of ordinary differential equations

$$\partial_t u_h - L_h u_h = f_h \ , \ t \in [0,T] \ , \tag{2.7}$$

where $L_h:V_h \to V_h$ in some sense approximates L and f_h approximates f. We shall normally assume that $u_h(0) = r_h u^0$.

We shall consider only those <u>fully-discrete</u> approximations which are defined on time-levels $0 = t_0 < t_1 < \ldots < t_n < \ldots < t_N = T$ and shall denote by a superscript n both the value of $u(t)$ at t_n and its approximation u_h^n, satisfying a one-step procedure

$$u_h^{n+1} = E_h^{(n)} u_h^n \ . \tag{2.8}$$

Here $E_h^{(n)}:V_h \to V_h$ approximates the evolution operator $\mathcal{E}(t_{n+1}-t_n)$ and again we shall normally take $u_h^0 = r_h u^0$. Multi-step schemes may be included in the formulation as follows: multiple sets of parameters, which may be identifiable with intermediate time levels, are included in the specification of V_h; prolongations p_h can still refer to a single main time-level but the definitions of restrictions r_h must be extended to families $\{u(t),\ 0 \leq t \leq T\}$ and the optimality definitions in (2.3) and (2.6) referred to $u(t) \in V$ for each t.

(a) Evolutionary error in the semi-discrete case

We suppose p_h and r_h to be time-independent. How they are chosen will depend on the discrete method and the analysis: a Galerkin method usually implies a prolongation and a difference method may imply a restriction; in either, a prolongation is implied if $p_h u_h$ is compared to u, and a restriction if u_h compared to $r_h u$. For simplicity of notation, we shall often denote $p_h u_h(t)$ by $U(t)$, and the parameters in $u_h(t)$ by $U_j(t)$. Then the projection $p_h r_h$ enables us to split the error as

$$u - U = (I - p_h r_h)u + p_h(r_h u - u_h) \quad . \tag{2.9}$$

The first term on the right is purely an approximation error in the space S_h and can be estimated from approximation theory. The second term, called the *evolutionary error*, is of greatest interest: we shall use the notation

$$e_h = r_h u - u_h \quad , \quad e = p_h e_h \quad . \tag{2.10}$$

We consider first the usual estimation procedure when u_h is generated by a Galerkin process and L is linear and coercive. Then, interpreting such expressions as $Lp_h u_h$ in a distributional sense, we decompose the difference between (2.1) and the prolongation of (2.7) as

$$\partial_t e = (p_h r_h - I)\partial_t u + L(u - p_h u_h) + (Lp_h - p_h L_h)u_h + (f - p_h f_h) \quad . \tag{2.11}$$

The first and last terms on the right are again approximation errors and the decomposition is aimed at isolating the middle terms. For a Galerkin process, $p_h \partial_t u_h$ is defined with $L_h = \tilde{r}_h L p_h$, $f_h = \tilde{r}_h f$ so that, for real u, f and L, we have

$$\langle (p_h L_h u_h + p_h f_h) - (Lp_h u_h + f), e \rangle = 0 \quad . \tag{2.12}$$

If, in addition, the restriction is to be defined such that $p_h r_h$ is the elliptic projection derived from L, we make the further splitting

$$L(u - p_h u_h) = L(u - p_h r_h u) + Le \tag{2.13}$$

and have

$$\langle L(u - p_h r_h u), e\rangle = 0 \quad . \tag{2.14}$$

Hence from (2.11) we obtain the energy estimate

$$\frac{d}{dt}\frac{1}{2}\|e\|^2 = \langle \partial_t e, e\rangle = \langle Le, e\rangle + \langle (p_h r_h - I)\partial_t u, e\rangle \tag{2.15}$$

giving

$$\frac{d}{dt}\|e\| + \alpha\|e\| \le \|(p_h r_h - I)\partial_t u\| \tag{2.16}$$

if $\langle Le, e\rangle \le -\alpha\|e\|^2$.

This decomposition is heavily oriented towards the Galerkin process and elliptic projection and the second term in (2.11) will often be difficult to estimate. It also has the disadvantage that the first term is present throughout and precludes the immediate study of superconvergence phenomena. We adopt instead a decomposition which places more emphasis on the discrete operator L_h. Thus we obtain from (2.7) and the restriction of (2.1)

$$\partial_t e_h - (L_h r_h u - L_h u_h) = (r_h L - L_h r_h)u + (r_h f - f_h) \quad , \tag{2.17}$$

reducing, when L and L_h are linear, to

$$\partial_t e_h - L_h e_h = (r_h L - L_h r_h)u + (r_h f - f_h) \quad . \tag{2.18}$$

The terms on the right in (2.17) and (2.18) we call the *truncation error* (T.E.). It is the key term in our analysis and may be estimated in a variety of ways. In particular, in the case treated above we have from the Galerkin process

$$\langle p_h L_h r_h u - L p_h r_h u, e\rangle = 0 \ , \ \langle p_h f_h - f, e\rangle = 0 \quad , \tag{2.19}$$

which gives the intermediate result

$$\langle p_h(\partial_t e_h - L_h e_h), e\rangle = \langle (p_h r_h L - L p_h r_h)u + (p_h r_h - I)f, e\rangle \quad . \tag{2.20}$$

Then, of course, when we introduce r_h so that $p_h r_h$ is the elliptic projector, we recover (2.15). But in general we shall prefer to retain (2.18) and concentrate attention on the difference of operators $r_h L - L_h r_h$: it is more general and more precise and can give pointwise bounds.

(b) Evolutionary error in the discrete case

For notational simplicity we consider the case where $t_{n+1} - t_n = \Delta t$, a constant, and denote by E_h^s the iterated operator E_h over s time steps. Then with the usual regrouping of terms we obtain from (2.2) and (2.8)

$$\begin{aligned} e_h^n &= r_h (E(\Delta t))^n u^0 - E_h^n r_h u^0 + (E_h^n r_h u^0 - E_h^n u_h^0) \\ &= [r_h E(\Delta t) u^{n-1} - E_h r_h u^{n-1}] + \ldots \\ &\ldots + [E_h^s r_h E(\Delta t) u^{n-s-1} - E_h^{s+1} r_h u^{n-s-1}] + \ldots \\ &\ldots + [E_h^{n-1} r_h E(\Delta t) u^0 - E_h^n r_h u^0] + (E_h^n r_h u^0 - E_h^n u_h^n) \quad . \end{aligned} \tag{2.21}$$

We assume that the approximate procedure is stable: that is, that there exist constants γ_0 and K such that

$$\|E_h^m v_h - E_h^m w_h\|_h \le K e^{\gamma_0 m \Delta t} \|v_h - w_h\|_h \, , \; \forall v_h, w_h \in V_h \quad . \tag{2.22}$$

Then we have

$$\begin{aligned} \|e_h^n\|_h &\le K \sum_{s=0}^{n-1} e^{\gamma_0 s \Delta t} \|[r_h E(\Delta t) - E_h r_h] u^{n-s-1}\|_h + K e^{\gamma_0 t_n} \| r_h u^0 - u_h \|_h \\ &\le K e^{\gamma_0 t_n} \{ \|r_h u^0 - u_h\|_h + t_n \sup_{0<t<t_n} \|(\Delta t)^{-1} [r_h E(\Delta t) - E_h r_h] u(t)\|_h \} \, , \end{aligned} \tag{2.23}$$

in a form similar to that in standard finite difference analysis (see Richtmyer and Morton [23]).

In most practical problems in several space dimensions, the error in the time discretisation is of much less concern than that in the spatial variables. It is then appropriate to estimate errors from (2.18) rather than from (2.23), which is also simpler. Note that if the time discretisation is unimportant and Euler's method is assumed, then $E_h u_h = u_h + \Delta t [L_h u_h + f_h]$: formally too, $E(\Delta t) u \rightarrow u + \Delta t [Lu + f]$ which then relates (2.23) to

(2.18). For conditions under which this second limit is valid the reader should consult the references on evolutionary equations already cited. Note, too, that the stability condition on the discrete scheme is replaced by the presence of the term $-L_h e_h$ on the left of (2.18).

Use of Fourier analysis

Fourier analysis gives the most precise insight into the local behavior of approximations and is therefore generally used in comparative assessment of methods on simple model problems - see, e.g., Swartz and Wendroff [29]. It has long been the most important tool in the study of the stability of difference schemes and, although subsequently rejected by Strang and Fix [27] in their analysis of the finite element method, their Fourier analysis of the method on a uniform mesh, given in an abreviated form by Strang [26], is very illuminating and forms the basis of the analysis in section 4. The difficulties arise when one wishes to be both precise and rigorous outside the simple case of linear, constant coefficient problems on a uniform mesh. However, both in differential equation theory and the stability theory of difference schemes, many of the results from Fourier analysis have been transferred to more general situations and much can be done here too.

In the simple case, and assuming too that the restriction operator is the same for all components of u, $\hat{r}_h$ the transform of r_h is a scalar multiplier and emerges from the homogeneous error term in (2.18) as a factor. Thus let k denote the d-component transform variable and $k\cdot x$ its inner product with the space variable x: then with $\hat{u}$, $\hat{L}$, etc. denoting transforms, we need to compute $\varepsilon(kh)$, where

$$\hat{L}_h/\hat{L} = 1 - \varepsilon(kh) \quad . \tag{2.24}$$

Equation (2.18), with f_h taken as $r_h f$, transforms to

$$\partial_t \hat{e}_h - \hat{L}_h \hat{e}_h = \hat{r}_h[\varepsilon(kh)\hat{L}(k)\hat{u}(k)] \quad , \tag{2.25}$$

and the possibility of superconvergence depends on $\varepsilon(kh)$. We use the term superconvergence as in Dupont [12] to refer to any property of u that is matched by the approximant to higher order than that characterising S_h. For evolutionary error it means the matching of $r_h u$ by u_h and will therefore normally depend on the choice of r_h: but in this simple case it does not.

As pointed out by Kreiss and Oliger [17], any discrete procedure needs at least two grid points per wave length to give minimal accuracy and in practice $\varepsilon(kh)$ will only be reasonably small for $|k| < \pi/2h$. It is therefore useful to split the truncation error by looking for bounds of the form $|\varepsilon(kh)| \leq C_1|kh|^\nu$ for $|kh| \leq \pi/2$, $|\varepsilon(kh)| \leq C_2$ for all kh, so as to obtain for $u = (2\pi)^{-d}\int e^{ik\cdot x}u(k)dk$

$$(2\pi)^d\ |T.E.| \leq C_1 h^\nu \int_{|k|\leq\pi/2h} |\hat{r}_h(k)|.|k^\nu\hat{L}(k)\hat{u}(k)|dk + C_2\int_{|k|\geq\pi/2h}|\hat{r}_h(k)\hat{L}(k)\hat{u}(k)|dk: \quad (2.26)$$

$\varepsilon(kh)$ is responsible for keeping the first term small and the second will depend on the spectrum $\hat{u}(k)$ of u and possible damping of inaccurate modes by $\hat{r}_h(k)$. *Then an error* bound for $\|e_h\|_h$ is obtained immediately from (2.18) and (2.26) with an assumption such as the semi-boundedness of L_h:$\mathrm{Re}\langle e_h, L_h e_h\rangle_h \geq -\alpha\|e_h\|_h^2$.

If now L is linear but the coefficients variable or the mesh non-uniform, u can still be resolved into its Fourier components but r_h cannot be taken through L_h and $\Sigma(kh)$ will depend strongly on its choice, now being defined for each node by

$$(L_h r_h e^{ik\cdot x})_j/(r_h L e^{ik\cdot x})_j = 1 - \varepsilon_j(kh) \quad . \qquad (2.27)$$

Error estimates can still be based on the equation

$$(\partial_t e_h - L_h e_h)_j = (2\pi)^{-d}\int u(k)\varepsilon_j(kh)(r_h L e^{ik\cdot x})_j dk \quad . \qquad (2.28)$$

We shall see in an example in section 7 how L_h and r_h may still be defined so as to retain superconvergence properties present in the uniform mesh case.

When L and L_h are non-linear, however, interaction between modes must be calculated and each case will have to be treated individually. Such studies have been carried out for various difference methods by a number of authors - see, e.g., Grammeltvedt [14]. We shall consider in sections 3 and 4, and again in section 7, the typical case when $L(u,v)$ is a product uv as in the advection operator $\underline{v}.\underline{\nabla}\underline{u}$.

3. APPLICATION TO A SIMPLE FINITE ELEMENT PROBLEM

We take the space V to be $L^2(-\infty,\infty)$ and consider the linear spline Galerkin method on a uniform mesh applied to $u_t = Lu$. An element of V_h consists of a set of nodal values $\{U_j\}$ at knots $x = jh$ and its prolongation p^1 is

$$U = \sum_{(j)} U_j \phi_j(x) \quad , \tag{3.1}$$

where summation is over $j \in Z$, $\phi_j(x) = \phi(x/h-j)$ and $\phi(s)$ is the piecewise linear shape function with $\phi(0) = 1$ and $\phi(\ell) = 0$ for non-zero $\ell \in Z$. We will use the optimal restriction operator* r^1, which makes $p^1 r^1$ the orthogonal projector into the space spanned by $\{\phi_j\}$; that is, if M is the mass matrix defined by $M_{mj} = \langle \phi_m, \phi_j \rangle$ and $u^{(\phi)}$ the vector defined by $u_m^{(\phi)} = \langle \phi_m, u \rangle$, then

$$r^1 u = M^{-1} u^{(\phi)} \quad . \tag{3.2}$$

In this case, M has components $2h/3$ for $m = j$, $h/6$ for $|m-j| = 1$ and zero otherwise. We will denote by $\{Q_j\}$ the nodal values of $r^1 u$.

Now taking the mode $u = \hat{u}(k)e^{ikx}$, we have

$$\langle \phi_m, u \rangle = \hat{u}(k) \int_{-\infty}^{\infty} e^{ikx} \phi(x/h-m)\,dx = h\hat{u}(k)e^{im\xi}\hat{\phi}(-\xi) \tag{3.3}$$

where $\xi = kh$ and $\hat{\phi}$ is the Fourier transform of ϕ. Hence the Fourier transform of the projection equations is obtained as

$$\frac{1}{6}\,(Q_{m-1} + 4\,Q_m + Q_{m+1}) = \hat{u}e^{im\xi}\hat{\phi}(-\xi)$$

$$\text{i.e.,}\quad Q_m \equiv \hat{M}^{-1}\hat{u}^{(\phi)} = \hat{u}e^{im\xi}\, 3\hat{\phi}(-\xi)/(2 + \cos\xi) \equiv \hat{u}e^{im\xi}\alpha(\xi), \text{ say } . \tag{3.4}$$

A simple computation shows that $\hat{\phi}(\xi) = (\frac{1}{2}\xi)^{-2} \sin^2 \frac{1}{2}\xi$ and hence

$$\alpha(\xi) = \frac{6(1 - \cos\xi)}{\xi^2(2 + \cos\xi)} \sim 1 + \frac{\xi^2}{12} \quad \text{as} \quad \xi \to 0 \quad . \tag{3.5}$$

* We will distinguish particular prolongations by superscripts (thus p^1 for linear splines) and always denote the corresponding optimal restriction with the same superscript.

This confirms the second order approximation error of the linear elements for a fixed k, as $\alpha(\xi)$ is the ratio of the nodal values of $p^1 r^1 u$ to those of u.

Next let us consider the evolutionary error when $L \equiv \partial_x$. We have for the nodal values when $u = \hat{u}e^{ikx}$,

$$(r^1 Lu)_j = ik(r^1 u)_j = ik\hat{u}e^{ij\xi}\alpha(\xi) \quad . \tag{3.6}$$

The discrete Galerkin operator L_h equals $r^1 Lp^1$, or $M^{-1}K$ in terms of the stiffness matrix $K_{mj} = \langle \phi_m, L\phi_j \rangle$: under Fourier transformation it becomes $\hat{M}^{-1}\hat{K}$ where $\hat{K} = i \sin \xi$, so we have

$$(L_h r^1 u)_j = (r^1 Lp^1 r^1 u)_j = \hat{M}^{-1}\hat{K}(r^1 u); = \frac{3 i \sin \xi}{h(2 + \cos \xi)} \hat{u}e^{ij\xi}\alpha(\xi) \quad . \tag{3.7}$$

The ratio of the nodal values (3.6) and (3.7) gives then

$$\frac{(L_h r^1 u)_j}{(r^1 Lu)_j} = \frac{3 i \sin \xi}{i\xi (2 + \cos \xi)} \sim 1 - \frac{\xi^4}{180} \quad \text{as} \quad \xi \to 0 \quad , \tag{3.8}$$

so defining $\varepsilon(kh)$ in (2.24). Whatever discrete norm $\|\cdot\|_h$ is chosen, this shows that the evolutionary error is $O(h^4)$ for a given k mode. Moreover, any restriction operator could have been used in the error definition $r_h u - u_h$: $\alpha(\xi)$ in (3.4) is just the Fourier transform $\hat{r}^1$ of r^1 and would be changed to $\hat{r}_h$, but the same factor would appear in (3.6) and (3.7) and cancel in (3.8).

To obtain precise error bounds for a general u will however depend on the choice of r_h. For r^1, the ξ^{-2} in $\alpha(\xi)$ provides some damping for the higher modes. Putting this with (3.8), which is quite accurate up to $\xi = \pi/2$, we can show for this case ($L \equiv \partial_x$ and $r_h = r^1$)

$$2\pi|\text{T.E.}| < C_1 h^4 \int_{k\in I_0} |k^5 \hat{u}(k)|dk + C_2 \sum_{n=1}^{\infty} n^{-2} \int_{k\in I_n} |k\hat{u}(k)|dk \quad , \tag{3.9}$$

where $I_n = \{k | n\pi \leq 2|k|h \leq (n+1)\pi\}$ and $C_1 \approx 0{\cdot}0090$, $C_2 \approx 4{\cdot}9$.

The choice r^1 is even more important for a non-linear operator, so let us now consider $L(u,v) = -u\partial_x v$, with L_h again the Galerkin operator. The discrete equations $\underline{W} = L_h(\underline{U},\underline{V}) \equiv M^{-1}K(\underline{U},\underline{V})$ become

$$h(W_{j-1}+4W_j+W_{j+1}) + (2U_j+U_{j-1})(V_j-V_{j-1}) + (2U_j+U_{j+1})(V_{j+1}-V_j) = 0 \quad . \tag{3.10}$$

The interaction of Fourier modes is exhibited by putting $v = \hat{v}e^{ik'x}$ to give

$$[L_h(r^1u,r^1v)]_j = \frac{-i[\sin\xi'+4\sin\frac{1}{2}\xi'\cos\frac{1}{2}\xi\cos\frac{1}{2}(\xi+\xi')]}{h[2+\cos(\xi+\xi')]}\hat{u}\hat{v}e^{ij(\xi+\xi')}\alpha(\xi)\alpha(\xi') \ . \tag{3.11}$$

On the other hand

$$[r^1L(u,v)]_j = -i\,k'\,\hat{u}\hat{v}e^{ij(\xi+\xi')}\alpha(\xi+\xi') \tag{3.12}$$

and denoting the ratio (3.11) to (3.12) by $\gamma_G(\xi,\xi')$ we have

$$\gamma_G(\xi,\xi') \sim 1 - (2\xi^3\xi'-7\xi^2\xi'^2-8\xi\xi'^3-4\xi'^4)/720 \quad \text{as} \quad \xi,\xi' \to 0$$

$$= 1 - 17\xi^4/720 \ , \quad \text{when} \quad \xi = \xi' \ . \tag{3.13}$$

Thus the fourth order accuracy is retained through the non-linear interaction under the optimal projection.

Clearly γ_G contains a factor $\alpha(\xi)\alpha(\xi')/\alpha(\xi+\xi')$ which will differ for different choices of r_h in the definition of evolutionary error. In the present case, this factor behaves like $1 - \frac{1}{6}\xi\xi'$ and cancels a similar factor in (3.11) to fourth order. If, however, the restriction operator r^c corresponding to point collocation were used, $\alpha(\xi) \equiv 1$ so that no cancellation occurs, $\gamma_G \sim 1 + \frac{1}{6}\xi\xi'$ and for this definition the Galerkin procedure for the product is only second order accurate. This is the interpretation used by Swartz and Wendroff [29], who therefore advocate simple multiplication of grid-point values for evaluating a product: under r^c that operation is then exact of course. We shall return to this point again in section 7.

4. EVOLUTIONARY ERROR IN THE SPLINE-GALERKIN METHOD

In [32] and [33] Thomée and Wendroff analysed the precise behaviour of the semi-discrete Galerkin method based on B-splines of order μ, when applied to linear differential operators on the real line. In [33] they showed that, for the periodic problem $\partial_t u = Lu$ on $[0,1]$, an accuracy of $O(h^\nu)$ was obtained at equally-spaced mesh points when L was of order m with C^∞ coefficients and initial data, where $\nu = 2\mu - m$ for m even and $\nu = 2\mu - m + 1$ for m odd and $\mu \geq (m+2)/2$. Their analysis uses a quasi-interpolant so that effectively they take the restriction to be collocation at the mesh points while prolongation produces an expansion using these coefficients and basis functions which are linear combinations of shifted B-splines.

Working in L^2, we shall use here the optimal restriction operator and show that the product operation then has accuracy $O(h^{2\mu})$. Since it is clear from the above that ∂_x is approximated to this accuracy for any restriction, we shall then be able to show that quite general non-linear differential equations can be approximated to $O(h^{2\mu})$ by using a succession of spline-Galerkin projections. The analysis closely follows [32] with only minor changes in notation.

Let χ be the characteristic function of the interval $[-\frac{1}{2}, \frac{1}{2}]$. Then the B-splines of order μ on a uniform mesh with spacing h can be defined as $\phi_j(x) = \phi(x/h - j)$, where $\phi = \chi * \chi * \cdots * \chi$ (μ factors) and $*$ represents convolution. The Fourier transform of ϕ is

$$\hat{\phi}(\xi) = [\hat{\chi}(\xi)]^\mu = [(\tfrac{1}{2}\xi)^{-1} \sin \tfrac{1}{2}\xi]^\mu \quad , \qquad (4.1)$$

and Thomée [32] introduces the trigonometric polynomials

$$g_{\mu,\sigma}(\xi) = (-i)^{\sigma-2\nu} h^{\sigma-1} \sum_{(m)} e^{-im\xi} \int_{-\infty}^{\infty} \partial_x^{\sigma-\nu} \phi_0(x) \partial_x^\nu \phi_m(x) dx \quad , \qquad (4.2)$$

where $\nu = [\frac{1}{2}\sigma]$ and $[\frac{1}{2}(\sigma+1)] < \mu$, and summation is over $m \in Z$: he proves that

$$g_{\mu,\sigma}(\xi) = \sum_{(m)} (\xi + 2\pi m)^\sigma [\hat{\phi}(\xi + 2\pi m)]^2 \quad . \qquad (4.3)$$

In particular, note that $hg_{\mu,0}(\xi)$ is the Fourier transform $\hat{M}$ of the mass matrix formed from the ϕ_j. It is also easily seen that there is a constant K_1 such that

$$1 \le [\hat{\chi}(\xi)]^{-2\mu} g_{\mu,0}(\xi) \le 1 + K_1(\xi/\pi)^{2\mu} \quad \text{for } |\xi| \le \pi \ . \tag{4.4}$$

From these definitions, the nodal parameters Q_j obtained from projecting a Fourier component $\hat{u}e^{ikx}$ with $\xi = kh$ are given by

$$\int_{-\infty}^{\infty}\{\sum_{(j)} Q_j\phi_{j-m}(x) - \hat{u}e^{im\xi}e^{ikx}\}\,\phi_0(x)dx \ , \ \forall m \in Z \ , \tag{4.5}$$

from which we see that

$$Q_j = \hat{u}e^{ij\xi}\alpha(\xi) \ , \quad \text{where } \alpha(\xi) = \hat{\phi}(\xi)/g_{\mu,0}(\xi) \ . \tag{4.6}$$

Hence

$$[1 + K_1(\xi/\pi)^{2\mu}]^{-1} \le [\hat{\chi}(\xi)]^{\mu}\alpha(\xi) \le 1 \ , \quad \text{for } |\xi| \le \pi \ . \tag{4.7}$$

Also, if $L \equiv \partial_x$, then writing simply p for the prolongation in this space and $\tilde{r}$ for the optimal restriction, $L_h = \tilde{r}Lp$ with Fourier transform $\hat{K}/\hat{M}$, so that

$$(L_h\tilde{r}u)_j = (\hat{K}/\hat{M})(\tilde{r}u)_j = \hat{M}^{-1}(\tilde{r}u)_j\, e^{-im\xi}\sum_{(q)}\int\phi_m(x)\phi_q'(x)e^{iq\xi}dx$$

$$= ih^{-1}[g_{\mu,1}(\xi)/g_{\mu,0}(\xi)](\tilde{r}u)_j \ . \tag{4.8}$$

It is easily deduced from (4.3) that $g_{\mu,1}/g_{\mu,0} = \xi[1 + O(\xi^{2\mu})]$ as $\xi \to 0$ so that we have $O(h^{2\mu})$ accuracy for the Galerkin approximation to ∂_x.

Now let us consider the product operation.

Lemma 4.1. If $L(u,v) \equiv uv$ and $(L_h\tilde{r})(u,v)$ is defined as $\tilde{r}[(p\tilde{r}u)(p\tilde{r}v)]$, then for $u, v \in H^{2\mu}$

$$\tilde{r}L(u,v) - (L_h\tilde{r})(u,v) = O(h^{2\mu}) \ , \text{ as } h \to 0 \ . \tag{4.9}$$

Proof

Taking first single components $\hat{u}e^{ikx}, \hat{v}e^{ik'x}$ with $\xi = kh, \eta = k'h$, the nodal parameters R_j for $L_h\tilde{r}$ are given by

$$\int\sum_{(j)} R_j\phi_j(x)\phi_q(x)\,dx = \int\sum_{(m)} \hat{u}e^{im\xi}\alpha(\xi)\phi_m(x)\sum_{(n)} \hat{v}e^{in\eta}\alpha(\eta)\phi_n(x)\phi_q(x)\,dx \quad . \tag{4.10}$$

If we introduce the factor $\gamma(\xi,\eta)$ by putting

$$R_j = \hat{u}\hat{v}e^{ij(\xi+\eta)}\alpha(\xi)\alpha(\eta)\gamma(\xi,\eta) \quad , \tag{4.11}$$

we obtain

$$hg_{\mu,0}(\xi+\eta)\gamma(\xi,\eta) = \int\sum_{(m)} e^{im\xi}\phi_m(x)\sum_{(n)} e^{in\eta}\phi_n(x)\phi_0(x)\,dx \quad . \tag{4.12}$$

The triple product can be evaluated by Poisson's summation formula

$$\sum_{(m)} e^{im\theta}\phi_m(x) = \sum_{(j)} \hat{\phi}(\theta+2\pi j)e^{i(\theta+2\pi j)x/h} \tag{4.13}$$

to give

$$g_{\mu,0}(\xi+\eta)\gamma(\xi,\eta) = \int\sum_{(m)}\sum_{(n)} \hat{\phi}(\xi+2\pi m)\hat{\phi}(\eta+2\pi n)e^{i[\xi+\eta+2\pi(m+n)]s}\phi_0(sh)\,ds$$

$$= \sum_{(m)}\sum_{(n)} \hat{\phi}(\xi+2\pi m)\hat{\phi}(\eta+2\pi n)\hat{\phi}(\xi+\eta+2\pi[m+n]) \quad . \tag{4.14}$$

Clearly γ like $g_{\mu,0}$ is 2π-periodic in its arguments and from (4.1) we see that

$$1 \le [\hat{\chi}(\xi)\hat{\chi}(\eta)\hat{\chi}(\xi+\eta)]^{-\mu}g_{\mu,0}(\xi+\eta)\gamma(\xi,\eta) \le 1 + K_2[(\xi/\pi)^{2\mu} + (\eta/\pi)^{2\mu}] \tag{4.15}$$

for $|\xi|$, $|\eta| \le \pi$ and some constant K_2. Now $(\tilde{r}L)_j$ contains the factor $\alpha(\xi+\eta)$, so if we denote by $1 - \varepsilon(\xi,\eta)$ the ratio $(L_h\tilde{r})_j/(\tilde{r}L)_j$ we have from (4.6), (4.11) and (4.14)

$$\varepsilon(\xi,\eta) = 1 - \alpha(\xi)\alpha(\eta)[\hat{\chi}(\xi+\eta)]^{-\mu}g_{\mu,0}(\xi+\eta)\gamma(\xi,\eta) \quad . \tag{4.16}$$

From the bounds (4.7) and (4.15) it follows that

$$|\varepsilon(\xi,\eta)| \le K_3[(\xi/\pi)^{2\mu} + (\eta/\pi)^{2\mu}], \quad \text{for } |\xi|,|\eta| \le \pi \quad . \tag{4.17}$$

Integrating now over all components of u and v, and noting that both $||L(u,v)||$ and $||L(p\tilde{r}u,p\tilde{r}v)||$ are bounded by $||u||\ ||v||$, we have

$$\tilde{r}L(u,v) - L_h\tilde{r}(u,v) = \tilde{r}[uv - (p\tilde{r}u)(p\tilde{r}v)]$$

$$= \int dk'' \; \tilde{r} \int \hat{u}(k) e^{ikx} \hat{v}(k''-k) e^{i(k''-k)} \varepsilon(kh, k''h-kh)\,dk \quad . \tag{4.18}$$

The integral over $|k| > \pi/h$, $|k''-k| > \pi/h$ converges to zero as $h \to 0$ and inside this range, (4.17) gives the required result with the constant including a factor $\|u\|_{H^{2\mu}} \|v\|_{H^{2\mu}}$.

From this bound and the similar one for the operation of differentiation one can state the following

Theorem 4.1. *Suppose* Lu *has the form* $L^q L^{q-1} \ldots L^1 u$, *where each* L^i *consists either of differentiation* ∂_x *or multiplication by* u *or a* C^∞ *function of* x. *Let* $L_h = L_h^q L_h^{q-1} \ldots L_h^1$, *where each* $L_h^i = \tilde{r}L^i p$: *that is, if* $L^i v \equiv \partial_x v$ *then* $L_h^i \tilde{r}v \equiv \tilde{r}\partial_x p\tilde{r}v$ *and if* $L^i v \equiv uv$ *or* $f(x)v$ *then* $L_h^i \tilde{r}v \equiv \tilde{r}(p\tilde{r}u)(p\tilde{r}v)$ *or* $\tilde{r}(p\tilde{r}f)(p\tilde{r}v)$; p *and* $\tilde{r}$ *are prolongations and restrictions for splines of order* $\mu \geq 2$. *Then*

$$\tilde{r}Lu - L_h\tilde{r}u = O(h^2) \quad \text{as} \quad h \to \infty \quad . \tag{4.19}$$

The proof follows immediately from the lemma and the earlier bound for differentiation after the decomposition

$$\begin{aligned} \tilde{r}L - L_h\tilde{r} = (\tilde{r}L^q - L_h^q\tilde{r}) L^{q-1} \cdot\cdot L' + \cdots \\ \cdots + L_h^q \cdot\cdot L_h^{q-s+1} (\tilde{r}L^s - L_h^s\tilde{r}) L^{s-1} \cdot\cdot L^1 + \cdots \\ \cdots + L_h^q \cdot\cdot L_h^2 (\tilde{r}L^1 - L_h^1\tilde{r}) \quad . \end{aligned} \tag{4.20}$$

Thus the full order of accuracy is preserved no matter how many derivatives and products are taken. In particular, for the linear splines $\mu = 2$, the usual loss of two orders of accuracy as one goes from ∂_x to ∂_x^2 is avoided by carrying out a projection between the two derivatives: the resulting scheme is, of course, much less compact than the standard scheme and the fourth order accuracy in this case will normally be better preserved by 'half-lumping' the mass matrix. In section 7, however, we show how the intermediate projection is valuable in the advection operation $u\,\partial_x u$.

One final remark on these methods - the L^2 norm clearly plays an important part in the thinking and $\partial_t U$ is constructed as a best approximation in that norm. However, when the final set of nodal parameters u_h^n is obtained, an optimal prolongation could be sought with respect to another space. For example, suppose that with linear splines a u_h^n was obtained very close to $r^1 u^n$. Then the best estimate of any functional of u^n from this data would depend on the smoothness assumed in u^n. For instance, nodal values might well be smoothed out: this would be consistent with the observation from (3.5) that $|(r^1 u)_j / u_j| > 1$ for all Fourier modes with $0 < kh \leq \pi$. (Compare the viewpoint in the field of optimal recovery - see Golumb and Weinberger [13], Meinguet [18], Micchelli and Rivlin [19], and references therein.)

5. INTERPRETATION OF STANDARD DIFFERENCE METHODS

In a sense a finite difference scheme is not concerned with prolongation operators and remains entirely in the discrete space. However, the standard stability theory always supposes some embedding of the approximants into the solution space of the differential problem and Raviart [22] does this in a manner which is directly antecedent to Aubin's analysis. Moreover, the choice of the initial data implies the use of a restriction operator and the interpretation of the final results implies that of a prolongation operator.

The simplest and most convenient operators from a theoretical viewpoint are those used by Raviart: the spatial region is sub-divided into rectangular cells C_j and r^0u is defined as the vector $\underline{Q} = \{Q_j\}$ obtained from averaging u over each cell; the corresponding prolongation $p^0\underline{Q} = \sum_{(j)} Q_j\theta_j(x)$, where θ_j is the characteristic function of the cell C_j. Then p^0 and r^0 are mutually optimal in L^2: see Temam [31] for developments in other spaces. This viewpoint is often used in a loose way in direct modelling of conservation laws, especially in fluid dynamics. However, in constructing such difference schemes these cell averages have to be inter-related or related to values at intermediate points and it is then that inconsistencies in the approach are often revealed, as compared with Temam's rigorous development.

It is much commoner in practice to regard the numbers in a difference calculation as representing grid-point values of the unknowns - and this is usually how the initial values are chosen: that is, the restriction r^c is a collocation. It is interesting then to consider what the optimal prolongation p^c should be to recover maximum information at each time step. From (2.5) p^c must be interpolatory. Also, Aubin's results do not apply in L^2 because r^c is not defined as a continuous operator there: we need to work in an underlying Sobolev space H^p where, by Sobolev's lemma, $p > d/2$. Because p^c has to interpolate, the norm in (2.5) under which the infimum is taken could in fact be just the L^2 norm of the p^{th} derivatives. The solution to this problem is then well-known in the case of a finite interval of the real line to be given by natural spline interpolation (see de Boor [5], Schoenberg [24]) and more generally will involve generalised splines (see, e.g. Schultz and Varga [25]).

It is not being suggested here that such procedures should be used in practice but that these relationships should be borne in mind when, for instance, designing contour-plotting packages: when processing the output of finite difference procedures one would expect them to be interpolatory; but, as noted above, when processing that from finite element schemes we could expect some smoothing of nodal values.

Let us consider now very briefly how three of the most important classes of difference scheme fit into the present framework. The Crank-Nicolson, or more generally the θ-method, applied to $\partial_t u = Lu$ leads to an evolution operator

$$u_h^{n+1} = E_h u_h^n \equiv [I_h - \Delta t\theta L_h]^{-1}[I_h + \Delta t(1-\theta)L_h]u_h^n \quad (5.1)$$

where L_h is a central difference approximation to L and some mild restriction on Δt may be needed to solve for u_h^{n+1} when L_h is non-linear. Then

$$[I_h - \Delta t\theta L_h](\Delta t)^{-1}[r^c E(\Delta t) - E_h r^c]u^n$$

$$= (\Delta t)^{-1}\{[I_h - \Delta t\theta L_h]r^c u^{n+1} - [I_h + \Delta t(1-\theta)L_h]r^c u^n\} \quad (5.2)$$

is the usual expression for the truncation error in a finite difference analysis. Similarly a Lax-Wendroff scheme may use two difference approximations $L_h^{(1)}$, $L_h^{(2)}$ together with an averaging operator A to give

$$E_h \equiv I + \Delta t L_h^{(2)}[A + \tfrac{1}{2}\Delta t L_h^{(1)}] \quad , \quad (5.3)$$

and more general Runge-Kutta schemes may be expressed in the same way, all giving a definition of truncation error identical with the usual one.

Multi-level schemes, like the leap-frog, require a little more care. We define V_h to represent discrete approximations at two time levels $\frac{1}{2}\Delta t$ apart, with r_h defined by the procedure used to obtain approximations at $t=0$, $t=\frac{1}{2}\Delta t$ from the initial data. The two meshes may or may not be staggered and we may cover the generalised case in which the two time levels are updated differently by using two difference operators $L_h^{(1)}$, $L_h^{(2)}$. Denoting by $u_h^{(1)}$, $u_h^{(2)}$ the components of u_h at the two levels, we have

$$\begin{pmatrix} I_h^{(1)} & 0 \\ -\Delta t L_h^{(1)} & I_h^{(2)} \end{pmatrix} \begin{pmatrix} u_h^{(1)} \\ u_h^{(2)} \end{pmatrix}^{n+1} = \begin{pmatrix} I_h^{(1)} & \Delta t L_h^{(2)} \\ 0 & I_h^{(2)} \end{pmatrix} \begin{pmatrix} u_h^{(1)} \\ u_h^{(2)} \end{pmatrix}^{n} , \tag{5.4}$$

$$E_h u_h \equiv \begin{pmatrix} I_h^{(1)} & \Delta t L_h^{(2)} \\ \Delta t L_h^{(1)} & I_h^{(2)} + (\Delta t)^2 L_h^{(1)} L_h^{(2)} \end{pmatrix} \begin{pmatrix} u_h^{(1)} \\ u_h^{(2)} \end{pmatrix} . \tag{5.5}$$

If L is linear and $-\mu^2$ is an eigenvalue of $L_h^{(1)} L_h^{(2)}$, then the eigenvalues of E_h are

$$\text{eig. } E_h = 1 \pm i\mu\Delta t \left[1 - (\tfrac{1}{2}\mu\Delta t)^2\right]^{\frac{1}{2}} - \tfrac{1}{2}\mu^2(\Delta t)^2 , \tag{5.6}$$

the upper sign giving the approximating modes and the lower the familiar spurious modes, when all variables are held at each level. The error $(\Delta t)^{-1}[r_h u^{n+1} - E_h r_h u^n]$ is then best studied by resolving $r_h u$ into the two sets of modes: for example, when u is scalar and $L_h^{(1)} = L_h^{(2)}$, one gets

$$u_h^{(2)} : u_h^{(1)} = \pm\left[1 - (\tfrac{1}{2}\mu\Delta t)^2\right]^{\frac{1}{2}} + \tfrac{1}{2} i\mu\Delta t \quad ; \tag{5.7}$$

that is, the true mode represents $u_h^{(2)}$ half a time-step ahead of $u_h^{(1)}$ and the spurious mode having $u_h^{(1)}$ and $u_h^{(2)}$ almost equal and opposite in sign. As is well known, such a linearised analysis carried out for $\mathcal{L}u = -u\partial_x u$ shows the spurious mode to be not only travelling in the wrong direction but to be amplified when the true mode is damped and thus capable of generating a non-linear instability.

Each of these schemes may be "hybridised" by using, for instance, Galerkin methods to generate L_h. The wave equation, treated as a system by leap-frog, is a useful example. Unstaggered linear elements under the Galerkin procedure give, of course, the asymptotic error $(1/180)\xi^4$ of (3.8): if the two meshes are staggered this is reduced by a factor $-13/32$ but at a cost of reduced stability.

6. A PETROV-GALERKIN METHOD FOR $\partial_t u = a\partial_x u$

If u is given, the Galerkin approximation to Lu is optimal in the L^2 sense so that, in a semi-discrete method or an explicit difference method in time as $\Delta t \to 0$, the Galerkin equations give an optimal approximation to the time derivative or difference of u. But for an implicit difference method one has to approximate the solution to equations like $u^{n+1} - \theta\Delta t L u^{n+1} = g^n$. If L is not self-adjoint it is likely that the Galerkin equations will give a poor approximation to u^{n+1}. Thus experiments for similar equilibrium problems are being conducted to find appropriate test functions for a Petrov-Galerkin approach - see [2] and [15] and references therein.

Suppose test functions χ_i are used to give an approximation to $\partial_t u = Lu$ by

$$(U^{n+1} - \theta\Delta t L U^{n+1} - [U^n + (1-\theta)\Delta t L U^n], \chi_i) = 0 \quad . \tag{6.1}$$

Then the operator $(I - \theta\Delta t L)$ in the inner product defines a new restriction operator s_h for obtaining $\{U_j^{n+1}\}$. The evolution operator in (2.8) is given by

$$E_h = s_h[I + (1-\theta)\Delta t L]p_h \quad , \tag{6.2}$$

while the r_h occuring in the truncation error of (2.23) will normally be different. Even for the explicit case $\theta = 0$, it may be useful to use the Petrov-Galerkin method.

We consider here the case when Euler's method is to be used on $u_t = a\partial_x u$ with linear elements on a uniform mesh and p^1, r^1 defined as in section 3. Now it is well-known that Galerkin's method gives a scheme which is unstable unless $\Delta t = O(\Delta x)^2$. On the other hand, when a is constant the characteristics can be followed exactly and there need be no evolutionary error. So we consider whether there is a choice of χ_i which will achieve this result.

Lemma 6.1. *Suppose the basis functions* ϕ_j *are such that we can set*

$$\partial_x \phi_j(x) = \pi_{j-1}(x) - \pi_j(x) \quad , \quad \psi_i(x) = \phi_j(x) - \pi_j(x) \tag{6.3}$$

and $\pi_j(x) = \pi(x/h-j)$ *for some* $\pi(x)$. *Then the Euler-Petrov-Galerkin method will follow the characteristics of* $\partial_t u = a\partial_x u$ *for* $a\Delta t = h$ *iff*.

$$(\chi_i , \psi_j) = 0 \quad \forall\, i, j \quad . \tag{6.4}$$

Proof

Then if $U^n = \sum Q_j\phi_j$, we need $U^{n+1} = \sum Q_{j+1}\phi_j$ for $a\Delta t = h$. Thus the following should be an identity in Q_j or, equivalently, in $Q_{j+1} - Q_j$

$$\langle \Sigma(Q_{j+1}-Q_j)\phi_j - h\sum Q_j \partial_x \phi_j \ , \ \chi_i \rangle = 0 \quad , \quad \forall\ i \ .$$

The coefficient of $(Q_{j+1}-Q_j)$ is just $\psi_j(x)$ from (6.3) so the result follows immediately.

For linear basic functions ϕ_j^1 , $\partial_x\phi_j^1 = (\phi_{j-1}^0-\phi_j^0)/h$, where ϕ_j^0 is the characteristic function of the interval $[jh,(j+1)h]$, so that $\psi_j = \phi_j^1-\phi_j^0$. Moreover, ψ_j is itself the difference of two triangular-shaped functions, $\psi_j = s_{j-1}-s_j$ where $s_j(x) = x/h-j$ in $[jh,(j+1)h]$ and $s_j(x) = 0$ otherwise. Thus we need $\langle \chi_i \ , \ s_j \rangle$ to be constant, independent of j. Clearly if χ_m is to be of the form $\chi_m(x) = \chi(x/h-m)$ with χ of compact support, we need

$$\int_m^{m+1} t\chi(t)dt = 0 \quad , \ \forall\ m \ . \tag{6.5}$$

We choose

$$\chi(t) = 4-6t, \quad \text{for} \quad t \in [0,1]; \ \chi(t) = 0, \quad \text{otherwise} \ . \tag{6.6}$$

Then $\int\chi dt = 1$ and for constant $a > 0$ we take test functions which are a linear combination of χ_i and ϕ_i^1. We must first establish stability.

<u>Theorem 6.1</u>. <u>The Petrov-Galerkin scheme given by</u>

$$\langle U^{n+1}- (U^n + a\Delta t\partial_x U^n) \ , \ (1-\nu)\phi_i^1 + \nu\chi_i \rangle = 0 \ , \ \forall\ i \quad , \tag{6.7}$$

<u>is stable for constant</u> a <u>on a uniform mesh if</u>

$$0 \le a\Delta t/h \le \nu \le 1 \quad . \tag{6.8}$$

Proof

We shall need the following results which can be easily computed.

<u>Lemma 6.2</u>. <u>For the basic functions</u> ϕ_j^0 , ϕ_j^1 <u>and</u> χ_j <u>we have</u>:

$$\langle \chi_i,\phi_j^0 \rangle = \langle \chi_i,\phi_j^1 \rangle = \langle \phi_i^0,\phi_j^0 \rangle = h\delta_{ij} = \frac{1}{4}\langle \chi_i,\chi_j \rangle \quad . \tag{6.9}$$

Denote by Φ_i the test functions in (6.7), defining a restriction operator s_h. Then we form first $V = p^1 s_h(a\partial_x U^n)$, multiply equation (6.7) by the coefficient $(U_i^{n+1}+U_i^n+\Delta t V_i)$ and sum over i to get

$$\langle U^{n+1}-(U^n+a\Delta t\partial_x U^n),\ \Sigma_{(i)}(U_i^{n+1}+U_i^n+\Delta tV_i)\Phi_i\rangle = 0 \quad . \tag{6.10}$$

Let us denote by $\underline{U}$ the vector of nodal values $\{U_j\}$ and by $p^0\underline{U}$, $p^\Phi\underline{U}$, $p^\chi\underline{U}$ its prologations using basis functions $\{\phi_j^0\}$, $\{\Phi_j\}$, $\{\chi_j\}$ respectively: we also retain the notation U for $p^1\underline{U}$. Then (6.10) can be written as

$$\langle U^{n+1}-(U^n+a\Delta t\partial_x U^n),\ [(1-\nu)p^1+\nu p^\chi](\underline{U}^{n+1}+\underline{U}^n+\Delta t\underline{V})\rangle = 0 \tag{6.11}$$

and, from the lemma, the cross products between $\underline{U}^n$ and $\underline{U}^{n+1}$ are symmetric and therefore cancel to give a difference of squares. Furthermore, *from the construction of* V it follows that for any vector $\underline{W}$,

$$\langle V-a\partial_x U^n,\ p^\Phi\underline{W}\rangle = 0 \tag{6.12}$$

So (6.11) reduces to

$$[(1-\nu)\|U^{n+1}\|^2+\nu\|p^0\underline{U}^{n+1}\|^2] - [(1-\nu)\|U^n\|^2+\nu\|p^0\underline{U}^n\|^2]$$

$$= \Delta t\langle a\partial_x U^n,\ [(1-\nu)p^1+\nu p^\chi](2\underline{U}^n+\Delta t\underline{V})\rangle \quad . \tag{6.13}$$

To simplify the right-hand side, note that $\langle\partial_x U^n,\ p^1\underline{U}^n\rangle = 0$ and that

$$\|h\partial_x U^n+p^\chi\underline{U}^n\|^2 = \|\Sigma_{(j)}U_j^n(\phi_{j-1}^0-\phi_j^0+\chi_j)\|^2 = \|p^\chi\underline{U}^n\|^2 \ :$$

expanding the sum gives therefore

$$2\langle\partial_x U^n,\ p^\chi\underline{U}^n\rangle = -h\|\partial_x U^n\|^2 \quad . \tag{6.14}$$

For the final term, we have from (6.12)

$$\langle a\partial_x U^n,\ p^\Phi\underline{V}\rangle = \langle p^1\underline{V},\ p^\Phi\underline{V}\rangle = (1-\nu)\|p^1\underline{V}\|^2+\nu\|p^0\underline{V}\|^2 \quad ,$$

while from the Cauchy-Schwarz inequality

$$2\langle a\partial_x U^n,\ p^\Phi\underline{V}\rangle \le (1-\nu)\|p^1\underline{V}\|^2+\nu\|p^\chi\underline{V}\|^2+\|a\partial_x U^n\|^2 \quad ,$$

so *that together we have*

$$\langle a\partial_x U^n,\ p^\Phi\underline{V}\rangle \le \|a\partial_x U^n\|^2 \quad . \tag{6.15}$$

Putting (6.14) with (6.15) gives for the right-hand side of (6.13),

$$[(a\Delta t)^2 - \nu h a\Delta t]\,||\partial_x U^n||^2$$

so that stability follows for $\nu h \geq a\Delta t$.

Next, we can consider accuracy and have the following:

Theorem 6.2. The Petrov-Galerkin scheme given by (6.7) is second order accurate if

$\nu = a\Delta t/h =$ constant < 1 .

Proof

We can use Fourier analysis, putting $u = \hat{u}e^{ikx}$. Then writing $\mu = a\Delta t/h$ and with $\alpha(\xi)$ given by (3.5),

$$(r^1 E(\Delta t)u)_j = \hat{u}\alpha(\xi)e^{i\xi(j+\mu)} \quad . \tag{6.16}$$

The mass matrix for (6.7) has transform $\hat{M} = h[\frac{1}{3}(1-\nu)(2+\cos\xi) + \nu]$, and we have

$$(E_h r^1 u)_j = \hat{u}\alpha(\xi)e^{i\xi j}\beta(\xi)$$

where
$$\beta(\xi) = 1 + a\Delta t\hat{M}^{-1}[(1-\nu)i\sin\xi + \nu(e^{i\xi}-1)] \quad . \tag{6.17}$$

The essential factor in the truncation error is therefore given by

$$(\Delta t)^{-1}[e^{i\xi\mu} - \beta(\xi)] \sim (\Delta t)^{-1}[\frac{1}{2}\mu(\nu-\mu)\xi^2 + \frac{1}{6}i\mu(\nu-\mu^2)\xi^3 + \ldots] \tag{6.18}$$

$$\text{as} \quad \xi \to 0 \quad .$$

The scheme is thus first order for $\nu > \mu$, second order for $\mu = \nu < 1$ and exact for $\mu = \nu = 1$.

It may be noted that this scheme with $\nu = \mu$ closely resembles the Lax-Wendroff method for this equation (or any other S_β^α method). In fact, the terms in U^n in (6.7) are exactly the same and it is only the presence of the mass matrix which distinguish the schemes: it has the effect of marginally improving the accuracy, for the coefficient of $i\xi^3$ in (6.18) is $\mu^2(1-\mu)/6$, while for Lax-Wendroff it is $\mu(1-\mu^2)/3$.

7. APPROXIMATION OF THE ADVECTION OPERATOR

In the foregoing sections, we have identified three easily implemented approximations to $u\partial_x u$, all based on the Galerkin method with linear elements but distinguished by the way in which the product is formed. The definition of the methods is contained in the expression for the truncation errors, in a semi-discrete process and with restriction r_h, as follows: the single-stage Galerkin method (SSG) gives

$$r_h(u\partial_x u) - r^1[(p^1 r_h u)(\partial_x p^1 r_h u)] \quad ; \qquad (7.1)$$

the double-stage Galerkin method (DSG) gives

$$r_h(u\partial_x u) - r^1[(p^1 r_h u)(p^1 r^1 \partial_x p^1 r_h u) \quad ; \qquad (7.2)$$

and point multiplication Galerkin (PMG) gives

$$r_h(u\partial_x u) - (r_h u)(r^1 \partial_x p^1 r_h u) \quad . \qquad (7.3)$$

The natural choice for r_h in the first two cases is r^1, while r^c is probably most natural for the third. For the purpose of comparison then, we tabulate for each case under both choices the leading term in $\varepsilon(\xi,\xi)$ for a single mode as introduced for (4.16). The entry for PMG under r^c follows immediately

Restriction	r^1	r^c
SSG	$\frac{17}{720}\xi^4$	$-\frac{1}{6}\xi^2$
DSG	$-\frac{1}{240}\xi^4$	$-\frac{1}{6}\xi^2$
PMG	$\frac{1}{6}\xi^2$	$\frac{1}{180}\xi^4$

Table Leading terms in $\varepsilon(\xi,\xi)$ for the three methods SSG, DSG, PMG under the restrictions r^1 and r^c.

from (3.8) because point multiplication is exact under r^c and there is no interference between the processes because $r^c(u\partial_x u) \equiv (r^c u)(r^c \partial_x u)$. The DSG entry under r^1 is new and is also the smallest: comparing with (3.13), it results from a $\gamma(\xi,\xi')$ which behaves like $1 + (2\xi^3\xi' + 3\xi^2\xi'^2 + 2\xi\xi'^3 - 4\xi'^4)/720$.

For smooth functions and under the restriction r^1, we would expect therefore that DSG would be up to nearly six times better than SSG. This will depend on the spread of modes in u since the total truncation error at node; is of the form

$$\text{T.E.} = \int dk e^{ikjh} \alpha(kh) \int d\kappa \frac{1}{4}(k-\kappa)\, \hat{u}(\frac{1}{2}k + \frac{1}{2}\kappa)\hat{u}(\frac{1}{2}k - \frac{1}{2}\kappa)\ \varepsilon(\theta+\delta,\ \theta-\delta) \quad , \tag{7.4}$$

where we have put $\xi+\xi' = 2\theta = kh$ and $\xi-\xi' = 2\delta = \kappa h$. Thus we compare

$$\varepsilon_{SSG}(\theta+\delta,\ \theta-\delta) \sim (17\theta^4 - 36\ \theta^3\delta + 10\theta^2\ {}^2 + 4\theta\delta^3 + 5\delta^4)/720 \tag{7.5}$$

and
$$\varepsilon_{DSG}(\theta+\delta,\ \theta-\delta) \sim -\ (3\theta^4 + 16\theta^3\delta\ -\ 30\theta^2\delta^2 + 16\theta\delta^3 - 5\delta^4)/720\ : \tag{7.6}$$

for a normally distributed spectrum, $\hat{u}(k) \propto e^{-\gamma k^2}$, these expressions would be multiplied by $e^{-\gamma(k^2+\kappa^2)/2}$ before being integrated over κ (or δ), so that the leading terms in (7.5) and (7.6) would be dominant.

By contrast, under these conditions PMG would be quite uncompetitive. On the other hand, considering the error $r^c u - u_h$ under the restriction r^c in this case, we need to compare (7.5) and (7.6) with

$$\varepsilon_{PMG}(\theta+\delta,\ \theta-\delta) \sim 4(\theta^4 - 4\theta^3\delta + 6\theta^2\delta^2 - 4\theta\delta^2 + \delta^4)/720 \quad , \tag{7.7}$$

which shows it to be quite comparable to DSG. The choice between these methods must therefore come down to a choice of restriction. With r^c, the factor $\alpha(kh)$ in (7.4) is missing and this becomes significant when contributions from $kh+2n\pi$ for all n are combined *in the coefficient* to e^{ikx}. Thus suppose we write in each case

$$E(k) = \max|\varepsilon|\ .\ \int d\kappa \left|\frac{1}{4}(k-\kappa)\ \hat{u}(\frac{1}{2}k + \frac{1}{2}\kappa)\ \hat{u}(\frac{1}{2}k - \frac{1}{2}\kappa)\right| \quad . \tag{7.8}$$

Then the coefficient of e^{ikx} in the truncation error is bounded by

$$\Sigma_{(n)}\ \alpha(kh+2n\pi)E(k+2n\pi/h) \quad : \tag{7.9}$$

for PMG under r^c this is an undamped sum over the E values; but for DSG (or SSG) under r^1, $\alpha(\xi)$ is rapidly damped for $|\xi| > \pi$, being reduced by a factor nine by $\xi = 3\pi/2$ and behaving asymptotically like ξ^{-2}. Such modes beyond the resolution of the grid are continuously created by the product operation and result in the phenomenon of aliasing: a sharply damped α is essential for their suppression. For the spline-Galerkin methods of section 4, $\alpha(\xi) \sim \xi^{-\mu}$ and for spectral methods (Orszag [21]) $\alpha(\xi) = 1$ for $|\xi| \leq \pi$ and $\alpha(\xi) = 0$ for $|\xi| > \pi$.

One further factor, besides the behaviour of errors for small ξ and the suppression of aliased modes for large ξ, needs to be considered in comparing methods. That is the growth rate of errors as expressed by the second term on the left of (2.17) or the stability bound in (2.22). For SSG one will have the same energy conservation properties as for the differential problem and the error growth can also be analysed in a similar way: for example, in the model problem $\partial_t u + u\partial_x u = 0$, one has for two semi-discrete approximations U and V on the whole real line

$$\frac{d}{dt}\|U-v\|^2 = -\langle \partial_x U^2 - \partial_x v^2 , U-V\rangle = \langle U^2-V^2 , \partial_x(U-V)\rangle$$

$$= \frac{1}{2}\langle U+V, \partial_x(U-V)^2\rangle = -\frac{1}{2}\langle \partial_x(U+V), (U-V)^2\rangle \quad . \tag{7.10}$$

Thus the deviation grows only when the mean solution is "compressive". On the other hand, in obtaining the greater accuracy with DSG one has sacrificed energy conservation and this result on error growth. For energy conservation one has now, where Δ_+, δ are the forward and central difference operators,

$$\langle \partial_t U + U(p^1 r^1 \partial_x U), U\rangle = 0 \quad ,$$

$$\text{i.e.,} \quad \frac{d}{dt}\frac{1}{2}\|U\|^2 = \langle (1-p^1r^1)\partial_x U, U^2\rangle \quad , \tag{7.11}$$

$$\sim (1/72)\Sigma_{(j)}(\Delta_+ U_j)^2(\delta^2\Delta_+ U_j) \quad : \tag{7.12}$$

thus energy grows in a typical compressive wave-front at a rate which is $O(h^4)$. Similarly instead of (7.10) one has

$$\frac{d}{dt}\|U-V\|^2 + \frac{1}{2}\langle\partial_x(U+V),(U-V)^2\rangle = 2\langle U(1-p^1r^1)\partial_x U-V(1-p^1r^1)\partial_x V,\ U-V\rangle$$

$$\sim (1/36)\Sigma_{(j)}[(\Delta_+U_j)(\delta^2\Delta_+U_j)-(\Delta_+V_j)(\delta^2\Delta_+V_j)]\Delta_+(U_j-V_j) \quad , \qquad (7.13)$$

which is much less useful. However, in numerical comparisons with the shallow water equations on a two dimensional triangular grid, Cullen [3,4] has found DSG to be significantly more accurate than SSG and both much better than PMG: he also noted that the two simpler Galerkin processes in DSG could be inmplemented more efficiently than the one process in SSG.

Finally, let us briefly consider the effect of an irregular mesh. With linear elements, the Galerkin method gives for $\underline{V} = r^1\partial_x p^1\underline{U}$ the equations

$$(\frac{1}{6}V_{j-1}+\frac{1}{3}V_j)h_{j-} + (\frac{1}{3}V_j+\frac{1}{6}V_{j+1})h_{j+} = \frac{1}{2}(U_{j+1}-U_{j-1}) \quad , \qquad (7.14)$$

where h_{j-} and h_{j+} are the intervals to the left and right of the j^{th} node. Now suppose that a co-ordinate transformation were made so as to give a regular mesh in a new co-ordinate y and denote dx/dy by $g(y)$. Then if $g(y)$ is piecewise constant, p^1 *denotes the same prolongation in both* x- *and* y- *space and the same vector* $\underline{V}$ is generated from

$$\langle\phi_i^1\ ,\ gp^1\underline{V}\rangle = \langle\phi_i^1\ ,\ \partial_y p^1\underline{U}\rangle \quad . \qquad (7.15)$$

That is, if we define r_y to be the restriction carried out in y-space with the weight function $g(y)$ in a least squares procedure, $\underline{V} = r_y(g^{-1}\partial_y p^1\underline{U})$. Now we can carry out a Fourier analysis in y-space and with $u = \hat{u}e^{iky}$ consider the truncation error

$$r_h(g^{-1}(y)\partial_y)u - (r_y\ g^{-1}(y)\partial_y p^1)r_h u \quad . \qquad (7.16)$$

This is probably simplest when r_h is chosen to be r_y: then the ratio of the nodal values is $\xi^{-1}\sin\xi(\hat{M}^{-1}\underline{a})_j$, where $\xi = kh$ and h is the y-mesh interval; the tridiagonal matrix $\hat{M}$ and vector $\underline{a}$ are given by

$$\hat{M} = (\ldots,\frac{1}{6}g_{j-}e^{-i\xi},\frac{2}{3},\frac{1}{6}g_{j+}e^{i\xi},\ldots) \quad , \qquad (7.17)$$

$$a_j = [2 - (g_{j+} e^{i\xi} + g_{j-} e^{i\xi}) + i\xi(g_{j+} - g_{j-})]/2(1 - \cos \xi) \quad , \qquad (7.18)$$

where g_{j-}, g_{j+} are the values of g to the left and right of the j^{th} node. This is a direct generalisation of the regular mesh case and shows more readily how the order of accuracy is lost than an analysis on the x-mesh. As Thomée and Wendroff [33] suggest, one can of course recover the full fourth order accuracy by choosing a smooth transformation $g(y)$ and replacing both r_y and r_h by r^1 in y-space: the method is now changed from (7.14) and the evolutionary error is defined by a projection in y-space, i.e. one has actually changed co-ordinate systems and no longer has an irregular grid.

REFERENCES

1. J. P. Aubin, Approximation of Elliptic Boundary Value Problems, John Wiley & Sons (New York), 1972.

2. I. Christie, D. F. Griffiths, A. R. Mitchell & O. C. Zienkiewicz, "Finite element methods for second order differential equations with significant first derivatives", Int. J. Num. Meth. Eng., v. 10, 1976, pp. 1389-1396.

3. M. J. P. Cullen, "A finite element method for a non-linear initial-value problem", J. Inst. Maths. Applics., v. 13, 1974, pp.233-248.

4. M. J. P. Cullen, "Application of the Finite Element Method to Numerical Weather Prediction", Ph.D. Thesis, Univ. of Reading, 1975.

5. C. de Boor, "Best approximation properties of spline functions of odd degree", J. Math. Mech., v. 12, 1963, pp.747-749.

6. J. E. Dendy, Jr., "Two methods of Galerkin type achieving optimal L^2 rates of convergence for first order hyperbolics", SIAM J. Num. Anal. v. 11, 1974, pp. 637-653.

7. J. E. Dendy, Jr., "Analysis of some Galerkin schemes for the solution of non-linear time-dependent problems", SIAM J. Num. Anal. v. 12, 1975, pp. 541-565.

8. J. Douglas, Jr. & T. Dupont, "Galerkin methods for parabolic problems", SIAM J. Num. Anal., v. 4, 1970, pp. 575-626.

9. J. Douglas, Jr., T. Dupont & M. F. Wheeler, "A quasi-projection approximation method applied to Galerkin procedures for parabolic and hyperbolic equations", MRC Technical Summary Report #1465, 1975.

10. T. Dupont, "L^2 estimates for Galerkin methods for second crder hyperbolic equations", SIAM J. Num. Anal., v. 10, 1973, pp. 880-889.

11. T. Dupont, "Galerkin methods for first order hyperbolics: an example", SIAM J. Num. Anal., v. 10, 1973, pp. 890-899.

12. T. Dupont, "A unified theory of superconvergence for Galerkin methods for two-point boundary problems", SIAM J. Num. Anal., v. 13, 1976, pp. 362-368.

13. M. Golomb & H. F. Weinberger, "Optimal approximation and error bounds", Proc. Symp. on Numerical Approximation, R. E. Langer (ed.), Univ. of Wisconsin Press (Madison), 1959, pp. 117-190.

14. A. Grammeltvedt, "A survey of finite difference schemes for the primitive equations for a barotropic fluid", Mon. Weath. Rev., v. 97, 1969, pp. 384-404.

15. J. C. Heinrich, P. S. Huyakorn, A. R. Mitchell & O. C. Zienkiewicz, "An upwind finite element scheme for two-dimensional convective transport equation", Int. J. Num. Meth. Eng., v. 11, 1977, pp. 131-143.

16. T. Kato, "Linear and quasi-linear equations of evolution of hyperbolic type", Proc. CIME Summer School, Cortona, 1976.

17. H. O. Kreiss & J. Olyer, "Comparison of accurate methods for the integration of hyperbolic equations", Tellus v. XXIV, 1972, pp. 199-215.

18. J. Meinguet, "Optimal approximation and error bounds in semi-normed spaces", Numer. Math. v. 10, 1967, pp. 370-388.

19. C. A. Micchelli & T. J. Rivlin, "A survey of optimal recovery", Proc. Symp. on Optimal Estimation in Approximation Theory, C. A. Micchelli & T. J. Rivlin (eds.), Plenum Press (New York), 1976, pp. 1-54.

20. B. Noble, "Error analysis of collocation methods for solving Fredholm integral equations", Proc. Conf. on Topics in Numerical Analysis, J. J. H. Miller (ed.), Academic Press (London), 1973, pp. 211-232.

21. S. A. Orszag, "Numerical simulation of incompressible flows within simple boundaries. I Galerkin (spectral) representations", Stud. Appl. Math. v. 50, 1971, pp. 293-327.

22. P. A. Raviart, "Sur l'approximation de certaines equations d'évolution linéaries et non-linéaries", J. Math. Pures Appld., v. 46, 1967, pp. 11-183.

23. R. D. Richtmeyer & K. W. Morton, Difference Methods for Initial-Value Problems, John Wiley & Sons (New York), 1967.

24. I. J. Schoenberg, "On interpolation by spline functions and its minimal properties", Proc. Conf. on Approximation Theory, P. L. Butzer (ed.), Birkhäuser Verlag (Basel), 1963, pp. 109-129.

25. M. H. Schultz & R. S. Varga, "L-splines", Numer. Math., v. 10, 1967, pp. 345-369.

26. G. Strang, "The finite element method and approximation theory", Proc. Conf. on Numerical Solution of P.D.E.s II (Synspade 1970), B. Hubbard (ed.), Academic Press (New York), 1971, pp. 547-583.

27. G. Strang & G. J. Fix, An Analysis of the Finite Element Method, Prentice-Hall (New York), 1973.

28. B. Swartz & B. Wendroff, "Generalised finite difference schemes", Math. Comp., v. 23, 1969, pp. 37-50.

29. B. Swartz & B. Wendroff, "The relative efficiency of finite difference and finite element methods. I Hyperbolic systems and splines", SIAM J. Num. Anal., v. 11, 1974, pp. 979-993.

30. L. Tartar, "Evolution equations in infinite dimensions", Proc. Symp. on Dynamical Systems, Vol. 1, L. Cesari, J. K. Hale & J. P. LaSalle (eds.), Academic Press (New York), 1976, pp. 167-177.

31. R. Temam, Numerical Analysis, D. Reidel Publ. Co. (Dordreckt, Holland), 1970.

32. V. Thomée, "Spline Galerkin methods for initial-value problems with constant coefficients", Lect. Notes in Maths. No. 363, Springer-Verlag, 1973, pp. 164-175.

33. V. Thomée & B. Wendroff, "Convergence estimates for Galerkin methods for variable coefficient initial-value problems", SIAM J. Num. Anal., v. 11, 1974, pp. 1059-1068.

34. M. F. Wheeler, "A priori L^2 error estimates for Galerkin approximations to parabolic partial differential equations", SIAM J. Num. Anal., v. 10, 1973, pp. 723-759.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
TSR #1832 (14) MRC-TSR-1832

2. GOVT ACCESSION NO.

3. ~~RECIPIENT'S CATALOG~~ NUMBER
(9) Technical

4. TITLE (and Subtitle)
(6) ANALYSIS OF EVOLUTIONARY ERROR IN FINITE ELEMENT AND OTHER METHODS.

5. TYPE OF REPORT & PERIOD COVERED
Summary Report, ~~no specific reporting period~~

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
(10) M. J. P. Cullen & K. W. Morton

8. CONTRACT OR GRANT NUMBER(s)
(15) DAAG29-75-C-0024

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Mathematics Research Center, University of Wisconsin
610 Walnut Street
Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Work Unit Number 7 -
Numerical Analysis

11. CONTROLLING OFFICE NAME AND ADDRESS
See Item 18.

12. REPORT DATE
(11) Feb~~ruary 19~~78

13. NUMBER OF PAGES
~~33~~ (12) 34p.

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U.S. Army Research Office	Meteorological Office	University of Reading
P. O. Box 12211	Bracknell	Dept. of Mathematics
Research Triangle Park	England	England
North Carolina 27709		

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Error analysis, Finite difference methods, Finite element methods, Spline-Galerkin, Petrov-Galerkin, Prolongation and restriction operators, Superconvergence, Advection equation.

v dot del(v)

partial sub t = a(partial sub x of u)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Restriction and prolongation operators are used to provide a unified framework for the discussion of errors in approximating evolutionary equations. A generalized truncation error enables the spline-Galerkin method to be studied in detail and the accuracy of various treatments of non-linear terms (such as the advection operator $v \cdot \nabla v$) compared: it is shown how a multi-stage Galerkin process can give errors which are $O(h^{2\mu})$ for splines of order μ and quite general differential operators. A Petrov-Galerkin method is derived for $\partial_t = a\partial_x u$ which is accurate and stable.

h to the 2mu power

mu

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 221 200

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)